Alignement multilingue pour l'étude contrastive : outils et applications

Olivier Kraif

Alignement multilingue pour l'étude contrastive : outils et applications

Olivier Kraif

Olivier.Kraif@u-grenoble3.fr

Laboratoire de Linguistique et didactique des langues étrangères et maternelles (LIDILEM)

Université Stendhal Grenoble 3

Riassunto

I corpora multilingui sono oggi facilmente accessibili in varie lingue, in grande quantità. Si possono inoltre trovare molti strumenti dedicati al trattamento di questi corpora : applicazioni per l'allineamento, *concordancer*, strumenti per la marcatura e l'annotazione XML, che sono distribuiti a volte con licenze gratuite o GPL. Tradizionalmente usati dai traduttori professionisti (per la costruzione delle Memorie di traduzione), o dai specialisti del trattamento automatico della lingua (i corpora di testi tradotti entrano nella confezione di certi sistemi di traduzione automatica), questi strumenti e dati linguistici sono ancora poco sfruttati nel campo della linguistica dei corpora.

Pensiamo invece che costituiscono una risorsa molto interessante per lo studio empirico dei contrasti linguistici, tanto al livello lessicale quanto a quello della morfosintassi.

Nella prima parte dell'articolo, presentiamo le tecniche più usate, nello scopo di valutare i risultati attuali dell'allineamento automatico. Più specificamente, daremmo una descrizione dei metodi generici implementati nel software Alinea, di cui i risultati sono stati misurati durante la recente campagna di evacuazione Arcade 2. Ne vedremmo le funzionalità principali : l'allineamento delle frasi, l'estrazione di corrispondenze lessicali, la ricerca di concordanze usando richieste bilingui su testi eventualmente etichettati e lemmatizzati, e anche l'estrazione automatica di un glossario bilingue.

Infine daremmo due esempi di osservazione contrastiva basate su queste funzionalità, che permettono di ricercare e comparare espressioni o costruzioni linguistiche complesse non solo al livello morfosintattico, ma anche sul piano delle strutture semantiche, seguendo le reti di differenze e analogie intrecciate dalle relazioni di equivalenza traduzionale.

Référence

Kraif O. (2008) Alignement multilingue pour l'étude contrastive : outils et applications, in Marie Hédiard (a cura di) Linguistica dei corpora, Strumenti e applicazioni, Edizioni dell'Università degli Studi di Cassino, pp. 83-99

1 Introduction

Les premiers travaux concernant l'alignement multilingue remontent à la fin des années 80 (Key & Röscheisen 1988, 1993). Ces travaux montraient qu'il était possible, à partir d'un texte et de sa traduction, d'identifier de façon automatique des relations de correspondance entre des segments équivalents, à différents niveaux de granularité: paragraphes, phrases, mots. Ces techniques, souvent basées sur des méthodes statistiques s'appuyant sur des indices superficiels tels que les nombres, les noms propres, les emprunts, les mots apparentés ressemblants, les longueurs de phrases, permettaient ainsi de transformer un corpus de textes 'parallèles' (i.e. en relation d'équivalence traductionnelle) en 'bi-texte', pour reprendre un terme de Harris (1988), où la relation d'équivalence traductionnelle est rendue explicite à travers un réseau de correspondances entre les unités constituantes.

Vers la fin des années 1990, la première campagne d'évaluation Arcade (Véronis 1997) montrait qu'on pouvait obtenir automatiquement un alignement de bonne qualité au niveau des phrases, avec plus de 95 % d'alignements corrects, pour des langues génétiquement proches telles que le français et l'anglais. En 2006, presque dix ans plus tard, lors de la campagne Arcade 2, les mêmes techniques ont prouvé leur efficacité, au niveau phrastique, avec des couples de langues éloignées tant sur le plan génétique qu'alphabétique, comme le français avec l'arabe, le chinois, le farsi, le grec, le japonais ou le russe. La correction[1] moyenne pour ces couples de langues se situait aux alentours de 87 % pour le meilleur système (Chiao *et al*. 2006).

Parallèlement à la maturation de ces techniques, on a assisté, avec la diffusion rapide du Web, a un véritable 'big-bang textuel' qui a permis de faciliter l'accès à de nombreux textes traduits et qui a également ouvert un nouvel espace pour la publication de traductions, sur un mode collaboratif. Avec les organismes internationaux tels que l'ONU, l'OMS ou l'UE, qui publient rapports, comptes-rendus et décisions législatives, traduits avec précision et rigueur en plusieurs langues officielles (le corpus de l'*Acquis Communautaire*[2] compte désormais 20 langues!); avec les projets issus de la mouvance du Logiciel Libre, qui publient des documentations techniques et des traductions collaboratives en de très nombreuses langues[3] ; avec les collections de textes littéraires numérisés, traduits - comme dans le projet Carmel (El-Bèze *et al.* 2006)[4] - et librement diffusés[5], la quantité de traductions disponibles dépasse - et de loin - le milliard de mots et concerne bien d'autres paires de langues que l'anglais et le français. Aujourd'hui, tout internaute peut se constituer rapidement une collection importante de tels textes - originaux et traductions - que nous appellerons désormais 'corpus multilingues parallèles'.

Dans la perspective de la linguistique contemporaine, qui réserve une large part à l'observation empirique des corpus, il devient donc pressant de saisir cette double opportunité:

- d'une part, la diffusion des techniques informatiques facilitant ces observations ;
- d'autre part, l'émergence d'une masse de phénomènes bi-textuels autorisant la mise en œuvre d'observations qualitatives mais aussi quantitatives. L'originalité des bi-textes (ou multi-textes quand on aligne plus de deux langues) est de recéler des phénomènes qui débordent l'étude linguistique endogène: ces phénomènes intéressent la traductologie, en tant qu'étude d'une pratique communicationnelle, mais aussi la linguistique contrastive, en tant qu'étude des différences d'organisation entre codes linguistiques.

Après avoir esquissé l'état de l'art des techniques d'alignement phrastique, à travers l'exemple particulier du logiciel Alinea, nous montrerons qu'il est également possible d'extraire des correspondances au niveau le plus fin (unités lexicales, voire morphèmes).

Nous verrons ensuite comment des outils de recherche, basés sur des techniques de TAL (pour Traitement automatique des langues), permettent d'explorer un corpus de façon bilingue et de cibler dans chaque langue des constructions particulières.

Nous examinerons également de quelle manière l'aspect "massif" d'un corpus permet de faire émerger des régularités intéressantes sur le plan contrastif, par la mise en évidence d'équivalences mais aussi de différences entre les codes. Nous verrons enfin que ces contrastes, par aller-retour depuis et vers le code source, permettent de révéler certaines structurations sémantiques du lexique de la langue.

2 Etat de l'art

Il existe aujourd'hui de nombreux logiciels d'alignement automatique. Certains sont des produits commerciaux, comme *Trados WinAlign*, ou *Mindo* de Babeling, d'autres sont issus de la recherche, et distribués gratuitement, comme *K-vec++*, *Giza++, Plug aligner*, ou *Alinea*, avec pour certains des licences de type ‘logiciel libre’.

Nous décrirons ici plus en détail le logiciel Alinea, développé par nous, et distribué gratuitement. Ce logiciel a été évalué lors de la campagne d'évaluation ARCADE 2 (Chiao *et al*. 2006), ce qui permet d'avoir une estimation rigoureuse de ses performances.

2.1 Fonctionnement d'Alinea

Pour bien comprendre le principe de l'alignement phrastique, il est utile d'avoir une représentation géométrique du processus. Un alignement étant un ensemble de paires associant des phrases ou groupes de phrases, chacune de ces paires peut être représentée par une petite surface rectangulaire dans l'espace bidimensionnel du bi-texte (surface proportionnelle aux longueurs des deux segments alignés). Ces regroupements sont appelés ‘transitions’ et correspondent à différents cas de correspondance : 1-1, 1-2, 2-1, etc. L'enchaînement de ces transitions est appelé un ‘chemin’

d'alignement, qui serpente peu ou prou autour de la diagonale du bi-texte, comme le montre la figure 1. Le but des logiciels d'alignement est de trouver le 'meilleur chemin possible', à la fois complet (i.e. sans ignorer de zones alignables), exact (i.e. en respectant la relation d'équivalence entre segments appariés) et de grain fin (car il est moins intéressant d'aligner au niveau des chapitres ou des paragraphes qu'au niveau des phrases).

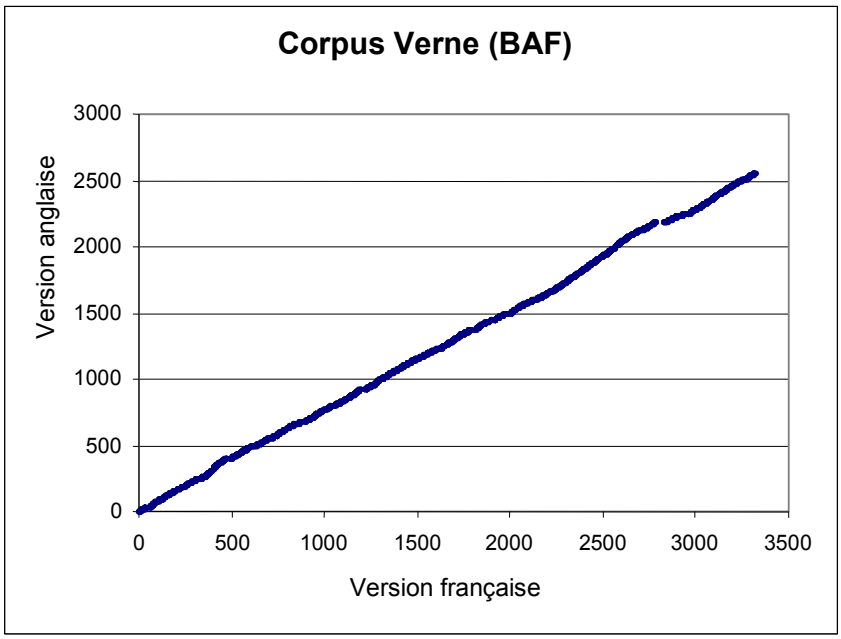

Figure 1 : chemin d'alignement extrait du corpus Verne (BAF)

Pour obtenir un tel alignement, *Alinea* procède en deux étapes:

- l'extraction de points d'ancrage**,** qui vise à confiner l'espace de recherche à l'intérieur d'îlots de confiance. Ces points d'ancrage doivent être extrêmement fiables, car ils sont déterminants pour la suite. Alinea peut se baser sur des points d'ancrage explicites (balises XML) ou extraire des points d'ancrage probables en s'appuyant sur des réseaux d'appariements de 'transfuges' concordants : nombres, noms propres, emprunts, etc. L'algorithme d'extraction est itératif et effectue un prédécoupage qui s'affine progressivement, en commençant par les transfuges les plus fiables (les nombres).

- l'alignement proprement dit, avec extraction du chemin complet optimal. On évalue la probabilité d'un chemin quelconque en fonction des indices disponibles : rapport des longueurs, appariement de chaînes identiques (transfuges), appariement de mots ressemblants (cognats), de mots possédant des distributions similaires ou de lexèmes équivalents. Un algorithme de programmation dynamique se charge alors d'extraire le chemin qui maximise cette probabilité. Pour optimiser les calculs, on se base en général sur un jeu de 8 transitions prédéfinies: 1-1, 1-0, 0-1, 2-1, 1-2, 2-2, 3-1, 1-3. Alinea permet cependant, lorsque les textes présentent des segmentations très divergentes, d'élargir la recherche à des transitions quelconques.

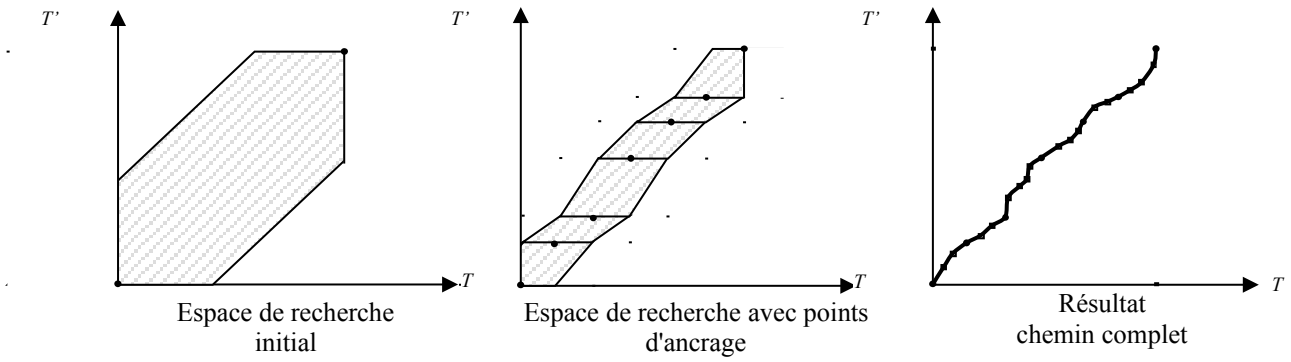

Figure 2 : réduction de l'espace de recherche avec des points d'ancrage

Notons que la pertinence relative des indices d'alignement dépend fortement de la typologie des textes et du couple de langues: entre l'anglais et le français, par exemple, la fréquence des cognats est très importante à l'intérieur de phrases alignées. Des textes riches en nombres seront également plus faciles à aligner. Dans l'exemple ci-dessous, ces indices sont très nombreux - il s'agit donc d'un cas de figure particulièrement favorable:

> - Le Comité préparatoire du cinquantième anniversaire de l 'Organisation des Nations Unies, que l ' Assemblée générale a créé par sa décision 46/472 du 13 avril 1992 , s 'est réuni cinq fois et s 'est mis d 'accord sur le choix d 'un thème. Crochet [*ACQUIS COMMUNAUTAIRE www*
>
> - The Preparatory Committee for the Fiftieth Anniversary of the United Nations, established by the General Assembly in decision 46/472 of 13 April 1992 , held five meetings.
> Agreement was reached by consensus on a theme for the anniversary. [*ACQUIS COMMUNAUTAIRE www*

Transfuges : *(Nations,Nations), (46/472,46/472), (13,13), (1992,1992)*
Cognats: *(Comité, Committee) (préparatoire, Preparatory) (anniversaire, Anniversary) (Nations, Nations) (Unies, United) (Assemblée, Assembly) (générale, General) (décision, decision) (avril, April) (thème, theme)*

Pour identifier les cognats, *Alinea* se base sur le calcul de la plus longue sous-chaîne commune (par exemple *préparatoire* et *preparatory* partagent une sous-chaîne de longueur 9 : p-r-p-a-r-a-t-o-r), ce qui ne nécessite aucune donnée linguistique. Quand on aligne des langues à alphabets différents, on ne peut plus s'appuyer aussi simplement sur les comparaisons de caractères: mais comme le montrent les résultats de la campagne Arcade 2, la plupart des traductions contiennent suffisamment de chaînes empruntées au système graphique de la langue source (nombres, sigles ou noms propres) pour qu'on puisse tirer parti de ces indices. Des prétraitements, tels que la translittération des nombres arabo-indiens, peuvent également être requis afin d'améliorer les résultats.
Enfin Alinea permet d'ajuster les paramètres concernant le rapport des longueurs et la pondération des différents indices, afin de s'adapter aux spécificités de chaque paires de langues. Lorsque les

indices de surface sont vraiment insuffisants, Alinea permet d'utiliser des ressources langagières, de type lexique bilingue, afin de compenser ce déficit d'information.

2.2 Résultats d'Alinea

Les premiers résultats de la campagne ARCADE II sont publiés dans Chiao *et al.* (2006). L'originalité de cette évaluation était de porter sur l'alignement du français avec d'une part des langues apparentées (anglais, allemand, espagnol, italien), et d'autre part des langues plus lointaines ou utilisant des alphabets différents (comme l'arabe, le chinois, le farsi, le grec, le japonais et le russe).

Pour le premier groupe, Alinea obtient des résultats corrects[6] à environ 98 %, à 3 dixièmes du meilleur système. Notons que les résultats sont meilleurs pour l'italien et l'espagnol que pour l'anglais et l'allemand, ce qui montre l'importance de la proximité génétique. Pour le second groupe, Alinea obtient les meilleurs résultats (mais seul un autre système était en compétition), avec une moyenne de 87,1 % : la dégradation des performances est avérée, mais pas catastrophique. Il existe tout un continuum entre les couples les plus propices (comme le français et le grec, avec 97,6 %) et les plus problématiques (comme le français et le japonais, avec seulement 78,9 % de correction). Notons que pour obtenir ces résultats, nous n'avons utilisé ni lexique bilingue, ni outil de translittération : le seul prétraitement était la segmentation en phrases.

Il existe une marge de progression réelle pour l'utilisateur qui prend le temps de régler ses paramétrages et d'enrichir Alinea d'un lexique bilingue (ce logiciel permettant aussi de constituer automatiquement ses propres ressources linguistiques). Quel que soit le couple de langues, on peut donc escompter des résultats compris dans une fourchette de 90 % à 99 % pour les techniques décrites ci-dessus, avec des traductions respectant les critères de parallélisme (sans omission ou ajout massifs).

3 Aligner au grain lexical

Avant d'aborder le niveau des mots, il faut noter que la notion d'alignement, bien qu'intuitive, est cependant plus complexe qu'il n'y paraît. L'exemple suivant a été extrait du corpus JOC[7]:

- Quelle action de planification de la logistique indispensable et de prévision budgétaire entreprend-elle en vue de la mise sur pied des autres programmes d' aide qui s' annoncent dès à présent ?

- What steps is the Commission taking to plan the necessary measures and ensure the availability of the necessary budgetary appropriations for further programmes of aid which already seem likely to be necessary ?

On peut se demander si les deux phrases, ainsi isolées, sont vraiment équivalentes. Le sens de "logistique" n'apparaît pas clairement dans la version anglaise et se réfère sans doute à des éléments

co-textuels. De même, la version anglaise insiste par de multiples répétitions sur l'aspect *necessary* des mesures à prendre. La construction du sens s'effectue au niveau textuel, et se manifeste par des macrostructures sémantiques caractérisant la cohésion (thématique, énonciative, stylistique, anaphorique, …) interne au texte, ainsi que sa cohérence vis-à-vis des référents extra-linguistiques[8]. En isolant une paire de phrases de son environnement co-textuel, on la prive du réseau de coréférences sémantiques sur lequel reposait en partie l'équivalence traductionnelle. De ce fait, il apparaît que la notion d'équivalence connaît un continuum de degrés. Initialement définie au niveau global, le texte traduit devant assumer de manière complète les fonctions communicatives que lui assigne le traducteur, dans le respect du sens de la source, l'équivalence traductionnelle devient de plus en plus lacunaire et morcelée quand on descend aux niveaux de granularité inférieurs. Inutile de dire qu'au niveau des mots (ou des morphèmes), elle vole en éclat, même si elle persiste ici et là pour certaines unités. Ainsi, comme nous l'avons montré par ailleurs (Kraif 2002), il y a une solution de continuité de la phrase au mot : une forte proportion des unités lexicales n'ont pas, en général, d'équivalent strict dans le texte traduit. En revanche on peut parler de correspondance lexicale : certaines unités conservent leur stabilité référentielle lors du passage à la traduction, et comme les « raisins dans la brioche », pour reprendre la métaphore de Seleskovitch citée par Laplace (1994), restent individualisables malgré la « chimie du sens » opérée par la traduction.

Or, ces correspondances peuvent être extraites de manière automatique, en se basant sur l'observation comparée des distributions des unités dans une vaste collection de textes alignés. En effet, connaissant les fréquence f_1 et f_2 de deux unités en langue source et en langue cible, il est possible de calculer leur fréquence de cooccurrence théorique F_{12} dans l'hypothèse où ces deux unités seraient indépendantes (cas où les cooccurrences seraient dues au hasard).

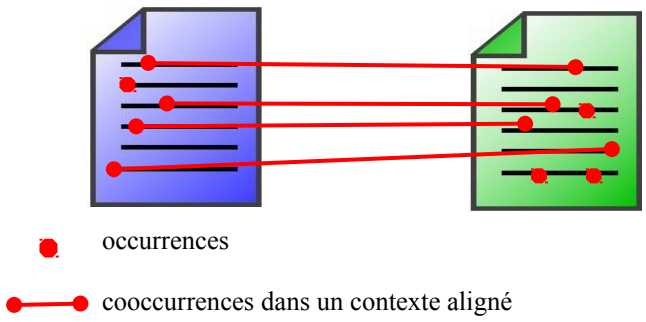

Figure 3 : Comptage des occurrences et cooccurrences : f_1= 5, f_2=7, f_{12}=4

Dans le cas d'unités équivalentes (p. ex. *anniversaire* et *anniversary*) la fréquence observée f_{12} est en général bien supérieure à la fréquence attendue F_{12}. Différents indices statistiques, tels que l'information mutuelle, le *t-score*, le rapport de vraisemblance, ou la probabilité de l'hypothèse nulle, permettent de mesurer ce degré d'association entre deux unités source et cible.

Pour Alinea, nous avons construit un indice mixte pour extraire, au sein de deux phrases alignées, les couples d'unités obtenant les meilleurs scores. En se basant sur les distributions, les ressemblances formelles, les positions dans les phrases alignées et l'identité des parties du discours, nous avons montré, dans Kraif & Chen (2004), qu'on pouvait obtenir des appariements corrects à 90% pour 88% des unités possédant des correspondances (en négligeant les mots vides: articles, prépositions, etc.), sur un corpus français-anglais (un extrait de *Madame Bovary*).

Les résultats de telles extractions dépendent avant tout de la taille du corpus d'apprentissage, duquel les statistiques d'occurrence et de cooccurrence ont été tirées: il faut compter au moins un million de mots dans chaque langue pour des résultats corrects à 80 %. Par ailleurs, la possibilité de faire correspondre des unités polylexicales (comme *à cause de*) ainsi que l'usage de corpus lemmatisés (Alinea peut traiter des corpus comportant lemmatisation et étiquetage morphosyntaxique[9]), permet d'obtenir une amélioration notable des résultats. Notons que les techniques employées permettent d'extraire des correspondances entre tout type d'unités préalablement identifiées dans les corpus : des mots composés, des expressions, mais aussi des traits morphosyntaxiques, tels que singulier/pluriel, masculin/féminin, présent/futur, etc. De la sorte, on peut étudier les correspondances sur un plan grammatical.

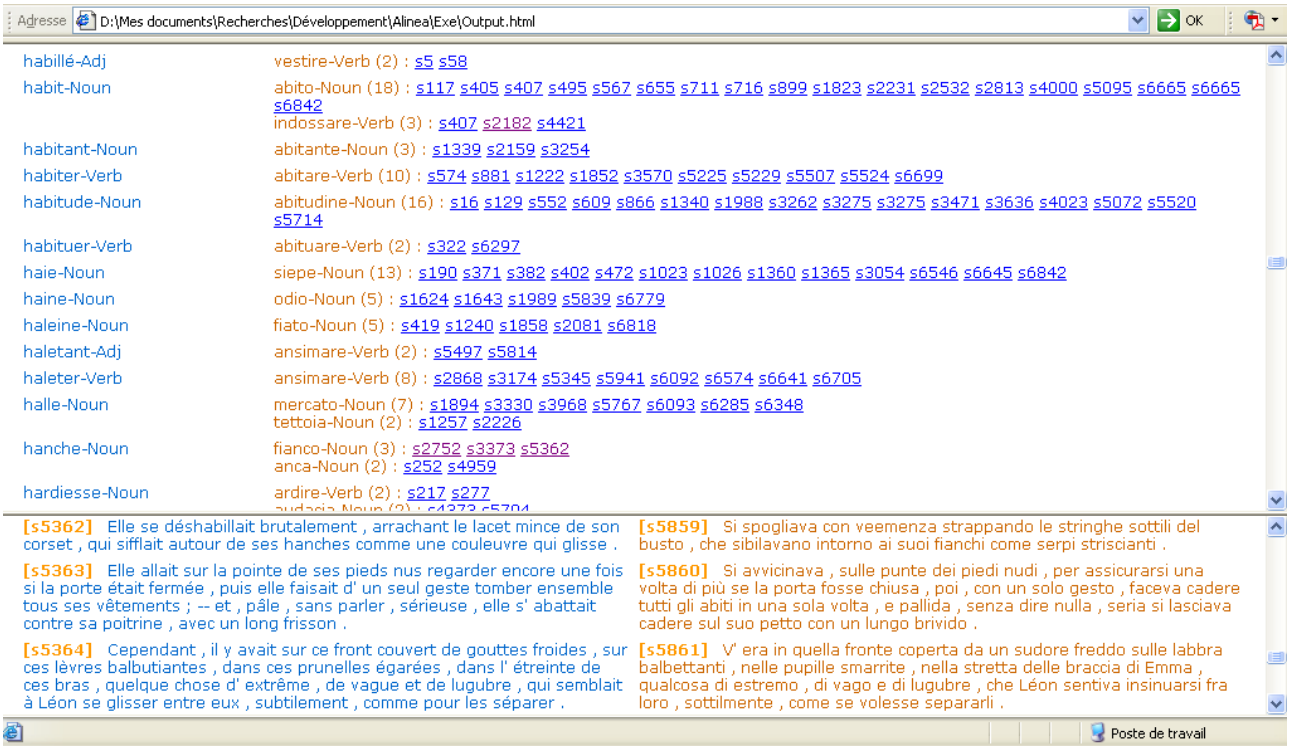

Figure 4 : Extrait d'un lexique bilingue tiré d'un alignement français-italien de *Madame Bovary*, de Flaubert

Autre résultat intéressant sur le plan contrastif, il est également possible d'extraire automatiquement un lexique bilingue spécifique à un corpus: pour filtrer le bruit et éliminer les correspondances trop liées à leur co-texte, il suffit de retenir les correspondances observées avec une fréquence statistiquement significative. Dans l'exemple de la figure 4, seules les correspondances observées plus de deux fois ont été retenues. Alinea permet d'exporter ce lexique sous format HTML, avec, pour chaque série d'associations lexicales, des hyperliens vers les phrases alignées du corpus. On peut ainsi comparer les exemples où *hanche* est traduit par *fianco* ou par *anca*.

4 Un outil pour l'étude contrastive

Au vu du lexique ainsi obtenu, on constate qu'il reste du bruit, souvent lié à des problèmes d'identification des unités polylexicales ou à des associations fortes entre unités cooccurrentes. Par exemple, *habit* est associé par erreur à *indossare*, à cause de constructions telles que *avoir un habit*, *porter un habit*, globalement traduites par *indossare*.

En augmentant la taille du corpus, on obtient des listes d'équivalences à la fois plus complètes et plus correctes. En effet, à mesure que les données deviennent statistiquement plus significatives, les régularités émergent et se distinguent des associations bruitées, plus instables par nature. Par ailleurs, les effets textuels, liés à l'idiosyncrasie d'un texte précis, à son sujet, aux habitudes de l'auteur, aux choix du traducteur, etc., s'estompent à mesure que le corpus augmente et devient plus représentatif de la langue générale (ou d'une langue de spécialité si l'on vise un corpus spécialisé).

Comme dans toute recherche de linguistique de corpus, on peut alors partir de l'observation du texte pour viser la langue. De ce point de vue, les bi-textes ne permettent pas seulement d'étudier deux langues, prises du point de vue du code, mais de les confronter et de les éclairer réciproquement, en s'appuyant sur les structures et les régularités originales que font apparaître les contrastes. Voici deux exemples de ces méthodes d'exploration contrastives.

4.1 Recherche d'expressions complexes

Il est possible d'aller plus loin que la recherche de formes simples ou d'expressions figées. En appliquant les outils de concordance à des corpus étiquetés et lemmatisés, comme Alinea le permet, il est possible de rechercher des expressions comportant un certain degré de variabilité morphosyntaxique (flexions, insertion de modifieurs, etc.) ou des constructions grammaticales précises. Par exemple, on peut étudier les différentes manières de traiter la concordance des temps entre principale et complétive, pour la modalité epistémique. Nous avons construit une requête bilingue autour des verbes *penser*, *croire*, et leurs équivalents italiens. Le langage d'Alinea permet de rechercher des lemmes (base) ou des traits morphosyntaxiques et d'utiliser des opérateurs d'expression régulières (tels que | + * ? !).

Requête :
Filtrage source : <base=/(croire|penser)><>?<base=que>
Filtrage cible : <base=/(credere|pensare)><>?<che><>?<>?
Résultats sur le corpus Bovary (échantillon):
fr310 Et quand je **pensais que** d' autres , à ce moment-là , étaient avec leurs bonnes petites femmes
it331 E quando **pensavo che** , in quello stesso momento , altri se ne stavano con le loro mogliettine
fr358 **Pensant qu'** après tout l'on ne risquait rien
it391 it392 **Pensando che** , in fin dei conti , non rischiava niente
fr560 elle ne pouvait s' imaginer à présent que ce calme où elle vivait fût le bonheur qu' elle avait rêvé .
it611 ; e adesso non riusciva a **credere che** la tranquillità nella quale viveva fosse davvero la felicità sognata .
fr1267 Elle ne **croyait pas que** les choses pussent se représenter les mêmes à des places différentes
it1395 Non **credeva possibile che** le cose potessero ripetersi nello stesso modo in luoghi diversi ,
fr1730 ils **croyaient que** c' était un sort .
it1907 **credevano** che fosse il malocchio .

Ces exemples illustrent la complexité des phénomènes de concordance, avec toutefois des régularités assez fortes: l'indicatif imparfait est utilisé dans les deux langues après *penser* et *pensare* à l'imparfait; le subjonctif imparfait est de mise après *credere* à l'imparfait, tandis que l'indicatif s'utilise après *croire*, sauf à la forme négative, qui renforce sémantiquement le doute. Dans ce dernier cas, on trouve bien un subjonctif imparfait, qui dénote l'ancienneté du texte, car on sait qu'il tombe aujourd'hui en obsolescence (alors qu'il est toujours usité en italien moderne).
Exploité avec de tels outils, le corpus révèle la richesse et la complexité des phénomènes visés, fournissant instantanément de nombreux exemples au linguiste.

4.2 Equivalence traductionnelle et classes sémantiques

De même que dans l'exemple précédent de *anca* et *fianco*, il y a fort à parier que des formes de la langue source qui partagent les mêmes équivalents, en langue cible, sont sémantiquement voisines. On peut donc utiliser la transitivité de la relation d'équivalence traductionnelle pour construire des classes de lexèmes synonymes ou sémantiquement proches.
Reprenons l'exemple donné ci-dessus: dans la phrase fr560, on trouve le verbe *s'imaginer* donné comme équivalent de *croire*. On peut alors relancer la requête sur cette classe élargie (*penser*, *croire*, *s'imaginer*). De nouveaux exemples donnent d'autres équivalents pour *s'imaginer*: *immaginare*, *supporre*, *aspettarsi*, *temere*, d'où une série d'autres équivalents français *suspecter*, *craindre* etc. Pour filtrer des correspondances peu fiables, on peut ne retenir que les équivalences apparaissant avec une certaine fréquence (par exemple au minimum deux fois). On continue ainsi "l'aller-retour", en relançant la recherche jusqu'à l'obtention de classes stables des deux côtés. Sur le corpus précédent, on obtient :

- penser, croire, imaginer, soupçonner, se douter, songer, se persuader, craindre, avoir peur, redouter (+ que/de)
- pensare, credere, immaginare, supporre, aspettarsi, temere, essere convinto, essere nella speranza, convincersi, sospettare (+ che/di), interpretare come

Certaines correspondances doivent être interprétées comme expressions polylexicales, pour avoir un sens. C'est le cas des sous-classes suivantes:

- porter à croire que, faire croire que, prétendre que
- indurre a ritenere di, lasciare ritenere che, far credere che

Toutes ces constructions permettent l'expression de la modalité subjective, à laquelle ils attribuent différentes valeurs: l'imagination, l'hypothèse, le doute, la certitude, la crainte. Ces classes, tout en étant assez larges, restent cohérentes et permettent d'explorer certaines dimensions sémantiques de la langue, sur une base purement empirique.

L'équivalence traductionnelle peut donc servir de révélateur: d'une part, des unités sémantiquement proches convergent vers les mêmes classes d'équivalents; d'autre part, une unité fortement polysémique est reliée à des classes relativement disjointes.

5 Conclusion

Les outils d'alignement et de concordance bilingue sont arrivés à maturité et permettent d'obtenir des résultats de bonne qualité avec peu d'intervention manuelle, et sans requérir une expertise en informatique ou en TAL. Mais certaines fonctionnalités, comme l'extraction automatique de lexiques bilingues, la construction de classes sémantiques, l'élaboration de requêtes tirant parti d'étiquetages morphosyntaxiques et de lemmatisation, sont encore relativement peu utilisées par la communauté des linguistes ou celle des traducteurs. On peut espérer que l'accès facilité à de grandes quantités de textes multilingues en ligne, ainsi que la disponibilité de certains outils qui franchissent le seuil des laboratoires, conférera à ces applications un développement rapide.

Mais au-delà des débouchés "naturels" de l'alignement en aide à la traduction, en lexicographie, en terminologie ou en didactique des langues, ces techniques permettront peut-être d'ouvrir un nouveau champ d'étude où le traitement automatique des langues, la traductologie et la linguistique de corpus convergeront vers l'étude contrastive. Grâce à l'observation de phénomènes statistiquement significatifs, à travers des corpus numériques de grande dimension, peut-être verra-t-on émerger des contrastes linguistiques invisibles "à l'œil nu".

6 Références

[Chiao *et al.* 2006], Chiao Y.-C., O. Kraif, D. Laurent, T. M. H. Nguyen, N. Semmar, F. Stuck, J. Véronis & W. Zaghouani, «Evaluation of multilingual text alignment systems: the ARCADE II project», in *Proceedings of the fifth International Conference on Language Resources and Evaluation, LREC* 2006, Genova, Italy.

[El-Bèze *et al.* 2006], El-Bèze M., Richard C., Meyer R., «Projet CARMEL : récits de voyages», in *Actes de TALN 2006*, Louvain-la-Neuve.

[Harris 1988], Harris B., «Are you Bi-Textual?», *Language Technology*, 7, 1988, 41-41.

[Kay *et al.* 1993], Kay M., Röscheisen, M. (1993) Text-Translation Alignement. *Computational Linguistics*, Morristown, NJ, vol. 19, n. 1,121-142

[Kraif 2002], Kraif O., «Translation alignment and lexical correspondences : a methodological reflection», in B. Altenberg & S. Granger (eds), Amsterdam, Benjamins Publisher, 2002, 271-290

[Kraif & Chen 2004], Kraif O., Chen B., «Combining clues for lexical level aligning using the Null hypothesis approach», in *Proceedings of Coling 2004*, Genève, August 2004, 1261-1264.

[Laplace 1994], Laplace C., *Théorie du langage et théorie de la traduction*, Paris, Didier érudition,1994.

[Rastier 1987], Rastier F., «Microsémantique et textualité», in M. Charolles, J.S. Petöfi, E. Sözer (eds), *Research in Text Connexity and Text Coherence*, Hambourg, Helmut Buske Verlag, 1987,147.

[Véronis 1997], Véronis J., «Une action d'évaluation des systèmes d'alignement de textes multilingues» ,in *1ères JST 1997 FRANCIL de l'AUPELF-UREF*, Avignon, 15-16 avril 1997, 191-197.

7 Outils d'alignement

- Alinea : http://www.u-grenoble3.fr/kraif
- Giza++ : http://www.isi.edu/~och/GIZA++.html
- K-vec++ : http://www.d.umn.edu/~tpederse/parallel.html
- Mindo : http://www.babeling.com/accueil.html
- Plug aligner : http://stp.ling.uu.se/~corpora/plug/pwa/
- Trados WinAlign : http://www.translation.net/trados_winalign.html

8 Notes

[1] Calculée en tant que *F-mesure*, moyenne harmonique de la précision et du rappel. La précision exprime la proportion d'alignements corrects par rapport aux alignements extraits (complémentaire du "bruit"). Le rappel exprime la proportion d'alignements corrects par rapport à l'alignement de référence (complémentaire du "silence").

[2]Le *ACQUIS COMMUNAUTAIRE Multilingual Corpus* est disponible en 20 langues à l'adresse : http://wt.jrc.it/lt/Acquis/ Il comporte environ 800 textes incluant l'ensemble des textes et des traités qui constituent le socle législatif de l'UE.

[3] Le corpus *Opus*, disponible à l'adresse http://logos.uio.no/opus/, contient des textes parallèles concernant jusqu'à 61 langues.

[4] Ce projet a permis de constituer un corpus de récits de voyage en quatre langues (français, anglais, italien, espagnol), alignés et comportant des annotations morphosyntaxiques, sémantiques et thématiques. Une partie du corpus est diffusé librement, à l'adresse : http://www.projetcarmel.org

[5] On trouve de nombreux textes sur le site du Projet *Gutenberg* (http://www.gutenberg.org/), sur le site de l'ABU (http://abu.cnam.fr/) et sur d'autres sites consacrés aux livres électroniques. *The Online Book Page* constitue un index assez complet des textes disponibles : http://onlinebooks.library.upenn.edu/ .

[6] La correction étant mesurée sur la base de la F-mesure, combinant précision et rappel.

[7] Le corpus JOC est composé de textes parallèles en neuf langues faisant partie du Journal Officiel de la Commission européenne (série C, année 1993). Les textes (au nombre de plusieurs milliers) sont constitués de questions écrites des parlementaires européens sur un large éventail de sujets, et des réponses correspondantes de la Commission européenne. La taille du corpus est d'environ 10,2 millions de mots, collectés et préparés dans le cadre des projets MLCC et MULTEXT. Le corpus est actuellement distribué sous licence par ELDA. Cf. l'adresse suivante : http://www.elda.org/catalogue/fr/text/W0017.html

[8] Ce que Rastier (1987) englobe simplement sous le terme de "textualité".

[9] L'étiquetage morphosyntaxique et la lemmatisation sont des techniques de base du traitement automatique des langues. L'étiquetage consiste à identifier les traits morphosyntaxiques (parties du discours, flexions, etc.) de formes apparaissant dans un texte. La lemmatisation consiste à identifier la forme canonique (le lemme) d'une forme fléchie (par exemple l'infinitif s'il s'agit d'un verbe). On arrive à obtenir automatiquement plus de 95% de correction sur des langues comme l'anglais ou le français.